CONTENTS

---

6   Data Reduction

6.1   Sufficient Statistics

Definition 1. Jointly Sufficient Statistics
Let $\mathbf{X} = (X_1, \ldots, X_n)$ have joint pdf $f(\mathbf{x}, \boldsymbol{\theta})$, and let $S = (S_1, \ldots, S_k)$ be a $k-$dimensional statistic. Then $S_1, \ldots, S_k$ is a set of **jointly sufficient statistics** for $\boldsymbol{\theta}$ if for any other vector of statistics, $\mathbf{T}$, the conditional pdf of $\mathbf{T}$ given $\mathbf{S} = \mathbf{s}$, denoted by $f_{\mathbf{T}|\mathbf{S}}(t)$, does not depend on $\boldsymbol{\theta}$. In the one-dimensional case, we simply say that S is a **sufficient statistic** for θ.

Example 1.
Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution. Show that $S = \sum X_i$ is a sufficient statistic for θ by definition.

Example 2.
Consider a random sample from an exponential distribution, $X_i \sim EXP(\theta)$. Show that $S = \sum X_i$ is a sufficient statistic for θ by definition.

6.2 Factorization Theorem

Theorem 1. Factorization Criterion

If $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots \ldots, x_n; \theta)$, and if $S = (S_1, \ldots, S_k)$, then $S_1, \ldots, S_k$ are jointly sufficient for θ if and only if

$$f(x_1, \ldots, x_k; \boldsymbol{\theta}) = g(\mathbf{s}; \boldsymbol{\theta}) h(x_1, \ldots, x_n)$$

where $g(\mathbf{s}; \boldsymbol{\theta})$ does not depend on $x_1, \ldots, x_n$, except through s, and $h(x_1, \ldots, x_n)$ does not involve θ.

Example 3.

Consider a random sample from a Gamma distribution, $X_i \sim Gamma(\alpha = 5, \theta)$. Show that $S = \sum X_i$ is a sufficient statistic for θ by factorization theorem.

Definition 2. If A is a set, then the indicator function of A, denoted by I_A is defined as

$$I_A = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}$$

Example 4. Consider a random sample from a uniform distribution, $X_i \sim U(0, \theta)$, where θ is unknown. Find a sufficient statistic for θ.

Example 5. Consider a random sample from a uniform distribution, $X_i \sim U(\theta, \theta + 1)$. Notice that the length of the interval is one unit, but the endpoints are assumed to be unknown. Show that $S_1 = X_{1:n}$ and $S_2 = X_{n:n}$ are jointly sufficient for θ.

Example 6.

Consider a random sample from a normal distribution, $X_i \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Show that $S_1 = \sum X_i$ and $S_2 = \sum X_i^2$ are jointly sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

6.3 Rao-Blackwell

Theorem 2. Rao-Blackwell Let $X_1, \ldots, X_n$, have joint pdf $f(x_1, \ldots, x_n; \boldsymbol{\theta})$, and let $S = (S_1, \ldots, S_k)$ be a vector of jointly sufficient statistics for $\boldsymbol{\theta}$. If T is any unbiased estimator of $\tau(\boldsymbol{\theta})$, and if $T^* = E(T|\mathbf{S})$, then

1. T^* is an unbiased estimator of $\tau(\boldsymbol{\theta})$

2. T^* is a function of $\mathbf{S}$, and

3. $V(T^*) \leq V(T)$ for every $\boldsymbol{\theta}$, and $V(T^*) < V(T)$ for some $\boldsymbol{\theta}$ unless $T^* = T$ with probability 1.

Example 7.

Let $X_1, X_2, \ldots, X_n$ be random sample of size n from a Poisson distribution with unknown mean λ. Let

$$T = \begin{cases} 1, & X_1 = x \\ 0, & \text{otherwise} \end{cases}.$$

Find $E\left[T \Big| \sum_{i=1}^{n} X_i = s\right]$.

Example 8.

Suppose $X_1, \ldots, X_{64}$ is a random sample from a normal distribution, $X_i \sim N(\mu, 25)$. Use the Rao-Blackwell theorem to find the UMVUE of $\nu = P[X \leq c]$.

6.4 Completeness

Definition 3. Completeness A family of density functions $\{f_{\mathbf{T}}(t; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Omega)\}$ is called complete if $E[u(T)] = 0$ for all $\boldsymbol{\theta} \in \Omega$ implies $u(T) = 0$ with probability 1 for all $\boldsymbol{\theta} \in \Omega$.

A sufficient statistic the density of which is a member of a complete family of density functions will be referred to as a **complete sufficient statistic**.

Example 9.
Let $X_1, \ldots, X_n$ denote a random sample from a Geometric distribution, $X_i \sim Geometric(p)$. Show that $S = \sum X_i$ is the complete sufficient statistic for p.

Example 10.
Let $X_1, \ldots, X_n$ be iid $N(\theta, a\theta^2)$, where a is known constant and $\theta > 0$. Show that the family of distribution is not complete.

Example 11.
Suppose that $X_1, \ldots, X_{43}$ is a random sample from a Geometric distribution, $X_i \sim \text{Geometric}(p)$. Let

$$T = \begin{cases} 1, & X_1 + \cdots + X_{10} = 10 \\ 0, & \text{otherwise} \end{cases}.$$

Find $E\left[T | \sum_{i=1}^{n} X_i = s\right]$.

6.5 Exponential Class

Definition 4. Exponeatial Class A density function is said to be a member of the regular exponential class if it can be expressed in the form

$$f(\mathbf{x}; \boldsymbol{\theta}) = c(\boldsymbol{\theta})h(\mathbf{x})e^{-\sum_{j=1}^{n} q_j(\boldsymbol{\theta})t_j(\mathbf{x})}, \mathbf{x} \in A$$

and zero otherwise, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a vector of k unknown parameters, if the parameter space has the form

$$\Omega = \{\boldsymbol{\theta} | a_i \leq \theta_i \leq b_i, i = 1, \ldots, k\}$$

(note that $a_i = -\infty$ and $b_i = \infty$ are permissible values), and if it satisfies regularity conditions 1, 2, and 3(a) or 3(b) given by:

1. The set $A = \{x : f(x; \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.

2. The functions $q_j(\theta)$ are nontrivial, functionally independent, continuous functions of the θ.

3.(a) For a continuous random variable, the derivatives $t_j(x)$ are linearly independent contin-

uous functions of x over A.

(b) For a discrete random variable, the $t_j(x)$ are nontrivial functions of x on A, and none is a linear function of the others.

For convenience, we will write that $f(x; \boldsymbol{\theta}))$ is a member of $REC(q_1, \ldots, q_k)$ or simply REC.

Theorem 3. $X_1, \ldots, X_n$, is a random sample from a member of the regular exponential class $REC(q_1, \ldots, q_k)$, then the statistics

$$S_1 = \sum_{i=1}^{n} t_1(X_i), \ldots, S_k = \sum_{i=1}^{n} t_k(X_i)$$

are a minimal set of complete sufficient statistics for $\theta_1, \ldots, \theta_k$.

Example 12.

Show that $X \sim Gamma(\alpha, \theta)$ belong to the regular exponential class, and use this information to find complete sufficient statistics based on a random sample $X_1, \ldots, X_n$.

6.6 Lehmann-Scheffe

Theorem 4. Lehmann-Scheffe Let $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n; \boldsymbol{\theta})$, and let S be a vector of jointly complete sufficient statistics for $\boldsymbol{\theta}$. If $T^* = t^*(\mathbf{S})$ is a statistic that is unbiased for $\tau(\boldsymbol{\theta})$ and a function of S, then T^* is a UMVUE of $\tau(\boldsymbol{\theta})$.

Example 13.

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution, $X_i \sim Bin(1, p)$; $0 < p < 1$. Find the UMVUE of p^2.

Example 14.

Suppose that $X_1, \ldots, X_n$ is a random sample from a Geometric distribution, $X_i \sim GEO(\theta)$,

(a) Show that the p.d.f. of X belongs to the regular exponential family.

(b) Based on the answer in a, find a complete and sufficient statistic for θ.

(c) Find the UMVUE of $\left[\frac{6\theta}{1-6(1-\theta)} \right]^n$.

Example 15.

Suppose that $X_1, \ldots, X_n$ is a random sample from a Poisson distribution, $X_i \sim POI(\theta)$. Find the UMVUE of $P(X = 0 \text{ or } 2) = (1 + \frac{1}{2}\theta^2)e^{-\theta}$ using Rao-Blackwell theorem.

6.7 Range-Dependent Exponential Class

Definition 5. A density function is said to be a member of the **range-dependent exponential class**, denoted by $RDEC(q_1, \ldots, q_k)$, if it satisfies regularity conditions 1 and 2 or 3 for $j = 3, 4, \ldots, k$ and if it has the form

$$f(x; \boldsymbol{\theta}) = c(\boldsymbol{\theta})h(x)e^{-\sum_{j=3}^{k} q_j(\theta_3, \ldots, \theta_k)t_j(x)}$$

where $A = \{x | q_1(\theta_1, \theta_2) < x < q_2(\theta_1, \theta_2)\}$ and $\boldsymbol{\theta} \in \Omega$.

1. The functions $q_j(\theta)$ are nontrivial, functionally independent, continuous functions of the θ.

2. For a continuous random variable, the derivatives $t_j(x)$ are linearly independent continuous functions of x over A.

3. For a discrete random variable, the $t_j(x)$ are nontrivial functions of x on A, and none is a linear function of the others.

For special cases:

1. The one-parameter case:
$$f(x; \theta) = c(\theta)h(x)$$
with
$$A = \{x | q_1(\theta), x < q_2(\theta)\}$$

2. The two-parameter case:
$$f(x; \theta_1, \theta_2) = c(\theta_1, \theta_2)h(x)$$
with
$$A = \{x | q_1(\theta_1, \theta_2), x < q_2(\theta_1, \theta_2)\}$$

Theorem 5.
Let $X_1, \ldots, X_n$ be a random sample from a member of the $RDEC(q_1, \ldots, qj)$.

1. If $k > 2$, then $S_1 = X_{1:n}$, $S_2 = X_{n:n}$ and $S_3, \ldots, S_k$ where $S_j = \sum_{i=1}^{n} t_j(X_i)$ are jointly sufficient for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

2. In the two-parameter case, $S_1 = X_{1:n}$, and $S_2 = X_{n:n}$ are jointly sufficient for $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

3. In the one-parameter case, $S = X_{1:n}$ and $S_2 = X_{n:n}$ are jointly sufficient for θ. If $q_1(\theta)$ is increasing and $q_2(\theta)$ is decreasing, then $T_1 = \min[q_1^{-1}(X_{1:n}), q_2^{-1}(X_{n:n})]$ is a single sufficient statistic for θ. If $q_1(\theta)$ is increasing and $q_2(\theta)$ is increasing, then $T_2 = \max[q_1^{-1}(X_{1:n}), q_2^{-1}(X_{n:n})]$ is a single sufficient statistic for θ.

If one of the limits is constant and the other depends on a single parameter, say θ, then the following theorem can be stated.

Theorem 6.
Suppose that $X_1, \ldots, X_n,,$ is a random sample from a member of the RDEC.

1. If $k > 2$ and the lower limit is constant, say $q_1(\theta) = a$, then $X_{n:n}$ and the statistics $\sum_{i=1}^{n} t_j(X_i)$ are jointly sufficient for θ and θ_j, $j = 3, \ldots, k$

2. If the upper limit is constant, say $q_2(\theta) = b$, then $X_{1:n}$ and the statistics $\sum_{i=1}^{n} t_j(X_i)$ are jointly sufficient for θ and θ_j; $j = 3, \ldots, k$.

3. In the one-parameter case, if $q_1(\theta)$ does not depend on θ, then $S_2 = X_{n:n}$ is sufficient for θ, and if $q_2(\theta)$ does not depend on θ, then $S_1 = X_{1:n}$ is sufficient for θ.

Example 16.
Consider the pdf
$$f(x; \theta) = \frac{1}{8\theta}, -3\theta < x < 5\theta$$
and zero otherwise. Find a single sufficient statistic for θ.

Example 17. Consider a two-parameter exponential distribution, $X \sim EXP(\theta, \eta)$. If $X_1, \ldots, X_n$ is a random sample, find the joint sufficient statistics for (θ, η).

Example 18.
Consider a random sample of size n from a uniform distribution, $X \sim U(\theta_1, \theta_2)$. Find the joint sufficient statistics for (θ_1, θ_2).

Example 19.
Consider a random sample of size n from a uniform distribution $X \sim U(0, \theta)$. Show that $S = X_{n:n}$ is a complete sufficient statistic for θ.

6.8 Basu Theorem

Theorem 7. Basu Let $X_1, \ldots, X_n$, have joint pdf $f(x_1, \ldots, x_n X, ; \boldsymbol{\theta})$; $\boldsymbol{\theta} \in \Omega$. Suppose that $S = (S_1, \ldots, S_k)$ where $S_1, \ldots, S_k$ are jointly complete sufficient statistics for θ, and suppose that T is any other statistic. If the distribution of T does not involve θ, then S and T are stochastically independent. In this case, T is call an ancillary statistic.

Example 20. Consider a random sample of size n from a normal distribution $X \sim N(\mu, \sigma^2)$. Use Basu Theorem to show that $\bar{X}$ and S^2 are independent.

Example 21.

Consider a random sample of size n from a two-parameter exponential distribution, $X_i \sim EXP(1, \eta)$.

1. Show that $T(X) = X_{1:n}$ is complete and sufficient for η.

2. Find the UMVUE of η.

3. Find the UMVUE of η^2.

4. Use Basu's Theorem to show that $X_{1:n}$ and $S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$ are independent.